Informatics Infrastructure at the start of the Second Decade of DNA Barcoding

SUJEEVAN RATNASINGHAM
BIODIVERSITY INSTITUTE OF ONTARIO
UNIVERSITY OF GUELPH
10+ Years
4M+ Records
0.5M+ Species
100+ Nations
Decade 1 - Capacity Building

- Data Standards
- Collaborative Networks
- Community Database
- Software Tools
Data Standards

- Quality
- Locus
- Provenance
- Complete

Consortium for the Barcode of Life

BOLD Systems

NCBI
Building the Library

The Barcode of Life Data Systems is designed to support the generation and application of DNA barcode data. The platform consists of four main modules: a data portal, a database of barcode clusters, an educational portal, and a data collection workbench.

**Sequence statistics**
- Barcode clusters for animals (99%): 380,042
- All Sequences: 3,207,920
- Barcode Sequences: 3,316,834

**Species coverage (formally described)**
- Animals: 143,471
- Plants: 50,000
- Fungi & Other Life: 16,206

Dataset: 3,941,206
Pre Data: 3,941,206

**Barcode Index Numbers (BINs)**
- A searchable database of Barcode Index Numbers (BINs), sequence clusters that closely approximate species.

**Workbench**
- An integrated data collection and analysis environment that securely supports the assembly and validation of DNA barcodes and ancillary sequences.

**Public Data Portal**
- A data retrieval interface that allows for searching over 17,000 public records in BOLD using multiple search criteria, including, but not limited to, geography, taxonomy, and deposition.

**DNA Barcode Education Portal**
- A custom platform for education and students to explore barcode data and contribute novel barcodes to the BOLD database.

**Barcodes**

**Linnean Species**

**BINs**

Timeline:
- 2004
- 2006
- 2008
- 2010
- 2012
- 2014

- 10,000,000
- 1,000,000
- 100,000
- 10,000
Collaborative Networks

- Registered Users ( Thousands )

- Data Sharing

- 2005 – 102 users from 30 institutions
1000+ Institutions from 94 countries sharing data on BOLD

BOLD User Network - 2015
Across Nations

Within Nations

100K+
10K – 100K
1K – 10K

BOLD User Network - 2015

66 Other Countries from Every Continent
Testing the Library Depth

Animals
- 400K+ Species
- 163K+ with Full Taxonomy
- 80% of the BOLD Library
- 500+ Orders

Test Data:
- 4000 species from 200 orders,
- 20 per order

BBC Tree of Life, 2014
Testing the Library Depth
Testing the Library Depth

Top Match Similarity

- >98%
- >95%
- >92%
- >90%

Years: 2005 to 2015
Testing the Library Depth

Amino Acid Composition

K-mers (k=3)

Ratnasingham, Ma, Hebert, in prep.
Barcode Index Number (BIN)

Algorithm

- Tuned to the marker (COX1)
- Fixed parameters for balanced OTU generation
- Uses prior threshold but refines for each group

Registry

- Occurrence of DNA Barcode (place and time)
- Aggregation of all associated metadata
- Reusable - works across studies
Oraesia Janzen 03  Notioplusia illustratata  Ctenoplusia sp. ANIC1
Importance of registering OTUs

- **Species**
- **Barcode Index Numbers**
- **Unregistered OTUs**

### LATITUDINAL RANGE

- **Fungi**
- **Plants**
- **Mollusca**
- **Araneae**
- **Fish**
- **Insecta**
- **Birds**
- **Mammalia**
The technology that links taxonomy and Star Trek

What made Star Trek's original tricorder a great piece of fictional technology, writes Maggie Koerth-Baker, wasn't its sci-fi looks. It was what it did.
Allow anyone, anywhere, to identify any organism.
lifescanner
LifeScanner Solution Overview

Sample Collection → Partner Sequencing Labs → Sequencing → ID Engine → Species Identification
YEP, THAT'S A PIGEON

Rise of the citizen scientist

From the oceans to the soil, technology is changing the part that amateurs can play in research. But this greater involvement raises concerns that must be addressed.
Moving into decade 2
Embrace Big Data

Complexity (Data volume & Dimensionality)

- What happened?
- Why it happened?
- What is happening?
- What might happen?

- Reporting
- Analysis
- Monitoring
- Forecasting

Impact
Support for multiple scales

$10^2$

$10^3$

$10^4$
Support for NGS based Barcoding
Launch of Beta on Sept 29, 2015

Advancing the Automation of Species Identification
Community Curated Libraries

- **Tier 1**
  - Purpose generated & reference specimens available
  - Barcode compliant & consistent
  - Key species (e.g. Dirty 22, Domesticated & Bush meat)

- **Tier 2**
  - Curated for consistency in taxonomic assignments
  - Barcode compliant & consistent
  - CITES/REDLIST (e.g. Endangered & controlled species)

- **Tier 3**
  - Mined from BOLD
  - Limited verification and only to be used as last resort
  - Disease vectors & invasive species

- **Disease vectors & invasive species**
  - 78%

- **CITES/REDLIST (e.g. Endangered & controlled species)**
  - 25%

- **Limited verification and only to be used as last resort**
  - 20%
More Analytical Tools
Community Developed Analytical Tools

BOLD4T

Plug-in Framework

6000+ Analytical Packages in CRAN
Analytical Pipelines

Simple Pipeline

1. Start
2. Select records of interest
3. Create a new dataset
4. BLAST each record
5. Add selected records to dataset
6. Within X% similarity
BOLD 4 + SAP Lumira
BOLD4 – Some other features

- Checklist Support (synonyms, progress, shopping lists)
- Data portal for core facilities
- Complete record histories
- RESL algorithm on your own datasets
- Storage and analysis of pre-clustered NGS data
Support for Metabarcoding

Support for the application of Next Generation Sequencing in DNA Barcoding Initiatives

From the Instrument to Interpretation

Storage of NGS Data
Support for storage of instrument data and experiment metadata in the cloud along with a publication pipeline that handles submission to GenBank SRA, & Dryad. Metadata templates are available for biodiversity surveys, ecological studies, and longitudinal studies. Metadata data standards are enforced in compliance with the Genomics Standards Consortium.

Data Preprocessing
Submitted NGS data files are preprocessed in the cloud based on project-specific parameters. Stream processing is employed with software that can be run on the desktop employing compression and bootstrapping to transfer data to BRAVE with minimal bandwidth impact. Computational tools employ FPGAs to achieve quick turnaround on analytical requests.

mBRAVE
Metabarcoding Research And Visualization Environment

Linkages & Partners

BOLD
SAP
Galaxy
ENA
NCBI
European Nucleotide Archive
Genomic Standards Consortium
Acknowledgements

BOLD Team

Paul Hebert
Dan Janzen & Winnie Hallwachs
Scott Miller
Axel Hausmann
Stefan Kremer
Many others!

Biodiversity Institute Collections Team
Canadian Center for DNA Barcoding Team

BOLD Users

[Logos of various organizations]